AD_____

Award Number: W81XWH-05-1-0267

TITLE: Functional Proteomic Analysis of Signaling Networks and Response to Targeted Therapy

PRINCIPAL INVESTIGATOR: Prahlad T Ram, Ph.D.

CONTRACTING ORGANIZATION: University of Texas
MD Anderson Cancer Center
Houston, Texas 77030

REPORT DATE: March 2007

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-03-2007	Annual	21 Feb 2006 – 20 Feb 2007

4. TITLE AND SUBTITLE

Functional Proteomic Analysis of Signaling Networks and Response to Targeted Therapy

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0267

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Prahlad T Ram, Ph.D.

E-Mail: pram@mdanderson.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas
MD Anderson Cancer Center
Houston, Texas 77030

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The purpose of the research done has been to determine the regulation of the EGFR network and identify how manipulations of the network alter signal flow to bypass targeted inhibitions. The scope of the project is to understand the network and determine which molecules have to be targeted to inhibit tumor cell proliferation. The major finding thus far are 1) We have performed the proteomic analysis of the signaling network in a panel of 4 breast cancer cell lines and determined the network response to EGF and targeted inhibitors. 2) We have determined how information flows within the network and feedback regulation. 3) Using these biological data we have developed a computational model and have made predictions to identify combinations of targets. We have completed the tasks listed for this time period and are on track as indicated in the original grant.

15. SUBJECT TERMS
No subject terms provided

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	20	**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

Page

Prahlad T. Ram
Functional proteomic analysis of signaling networks and response to targeted therapy
DOD-IDEA BC044268

Progress report year 2

Introduction

The purpose of the research done has been to determine the regulation of the EGFR network and identify how manipulations of the network alter signal flow to bypass targeted inhibitions. The scope of the project is to understand the network and determine which molecules have to be targeted to inhibit tumor cell proliferation. In the past year we have been very active in our research efforts. We have accomplished many of the tasks laid out in the SOW. Task 1A, 1B, 1D and 2A were completed during year 1 and were included in last years report. We have been working on Tasks 1C and 2B during the past funding year and the data from these two tasks are shown in this annual update.

Body

Task 1C of the proposal was to determine the changes in the EGFR signaling network in response to EGFR inhibitors, and task 2B was to determine the information flow within the network. We have accomplished both these tasks. We are currently working on Task 3.

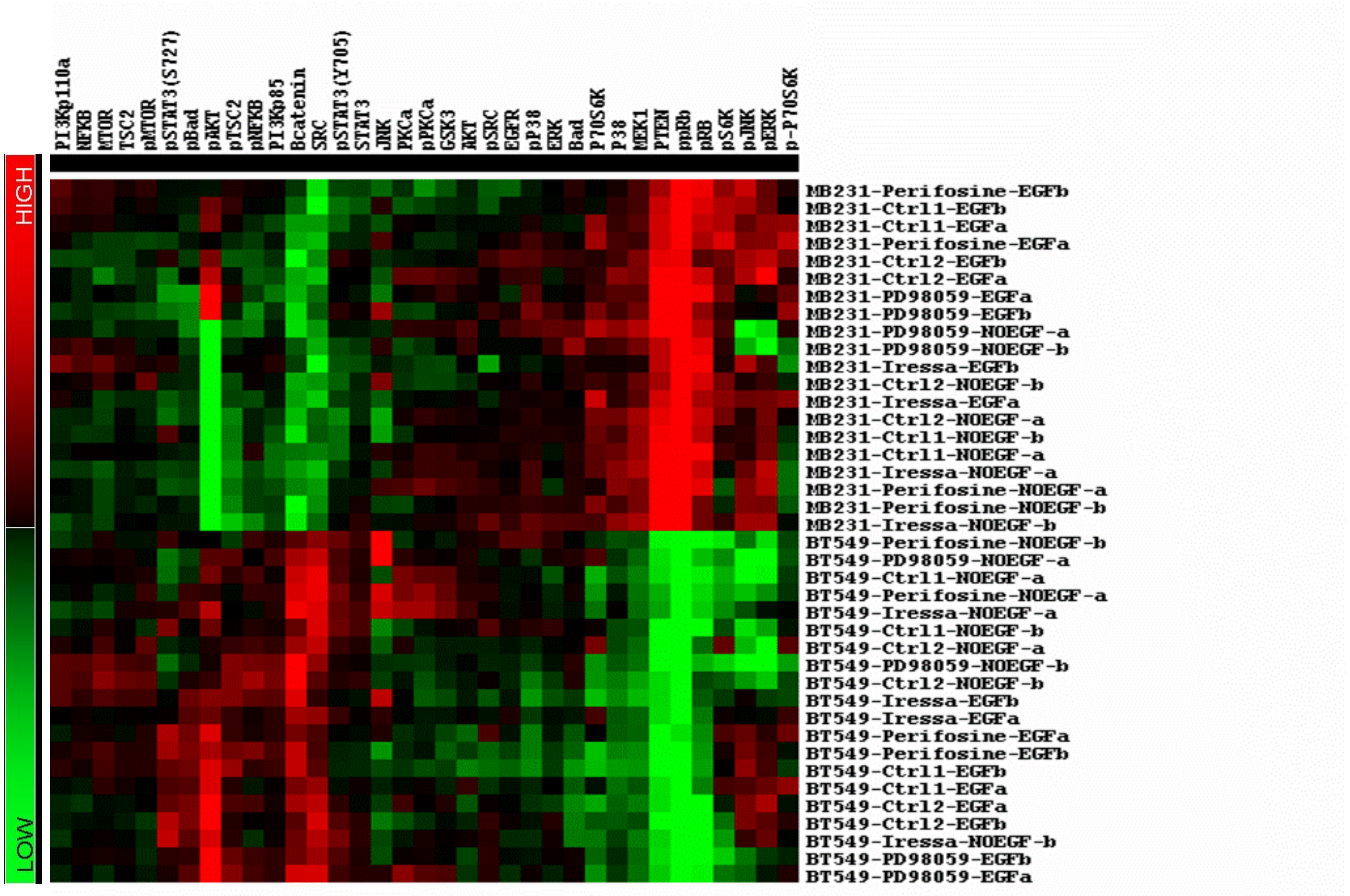

Figure 1. Reverse phase protein array data from BT549 and MDA-MB-231 breast tumor cells.

Shown in Figure 1 is the data from the reverse phase protein arrays for two breast tumor cell lines (BT549 and MDA-MB-231). The cells were serum starved overnight and treated with the EGFR inhibitor Iressa, or the MEK inhibitor PD98059 or the AKT inhibitor Perifosine for 2 hours. Control cells were treated with DMSO. The cells were then stimulated with EGF or vehicle for 30 minutes and the cells were then lysed. The soluble proteins were spotted onto the reverse phase protein arrays, printed on Fast20 nitrocellulose coated slides and probed with antibodies to the different phosphor and total proteins of the EGFR network. The data shows the quantitative changes and is visualized in the heat map, where red indicates high and green indicates low. The lysates were probed with 40 different antibodies. We see the expected changes in activity of signaling molecules in response to EGF and upon incubation with targeted inhibitors.

The same data is shown in Figures 2-4 as nodes and their interactions in the signaling network. Visualizing the data in the two different ways allows us to identify clusters of changes in signaling proteins (i.e. clustering heat maps Fig1) as well as in their context of their interactions within the network (Figs 2-4).

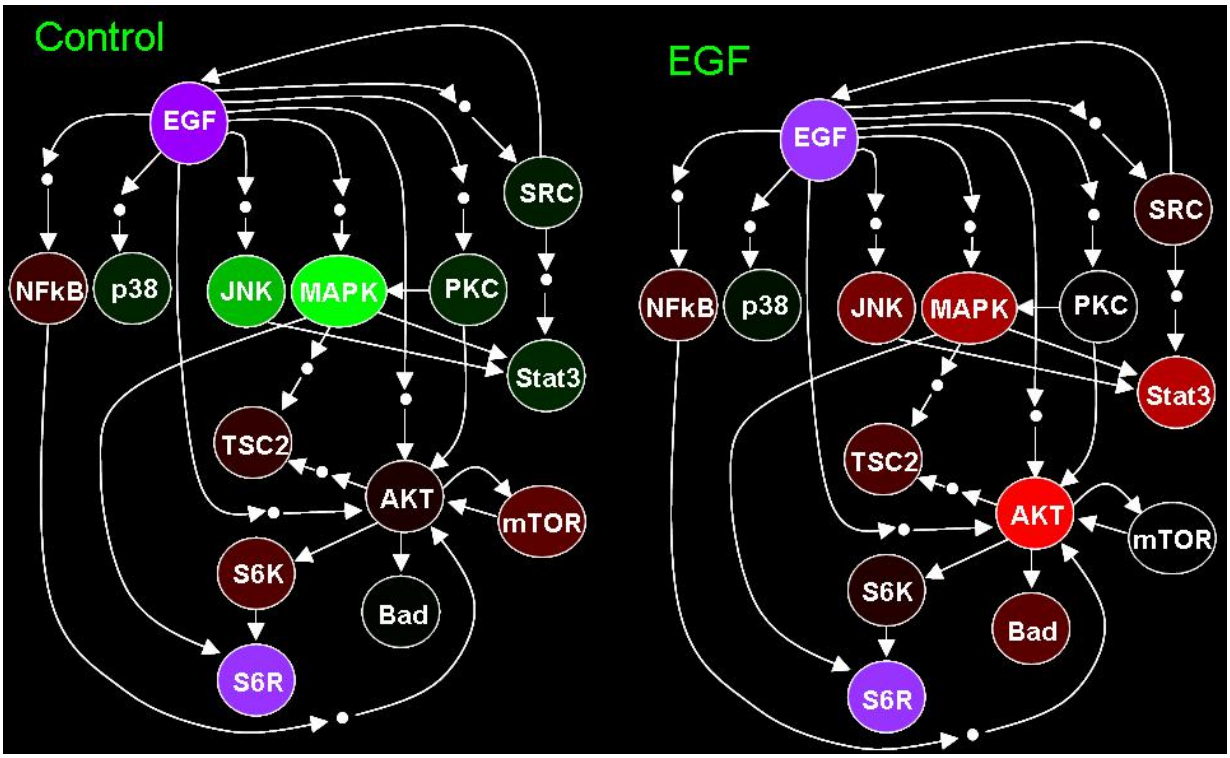

Figure 2. Network diagram showing information flow from the protein array data.

In Figure 2 we show signal flow through the network where the colors are indicative of levels of activity. Red signifies high activity and green low activity. We see on the left panel that in control cells MAPK, Stat3, Src and AKT activity is low. Upon stimulation with EGF there is an increase in activity of MAPK, JNK, AKT, Src and Stat3. These data have been incorporated into network model previously developed.

We next determined what changes occur in the network in response to EGFR inhibitors. Figure shows the data from the experiments where cells treated with the EGFR inhibitor Iressa for 2 hours and stimulated with EGF. The data shows that Iressa blocks the EGF activation of JNK, MAPK, and Stat3, and partially blocks the activation of AKT.

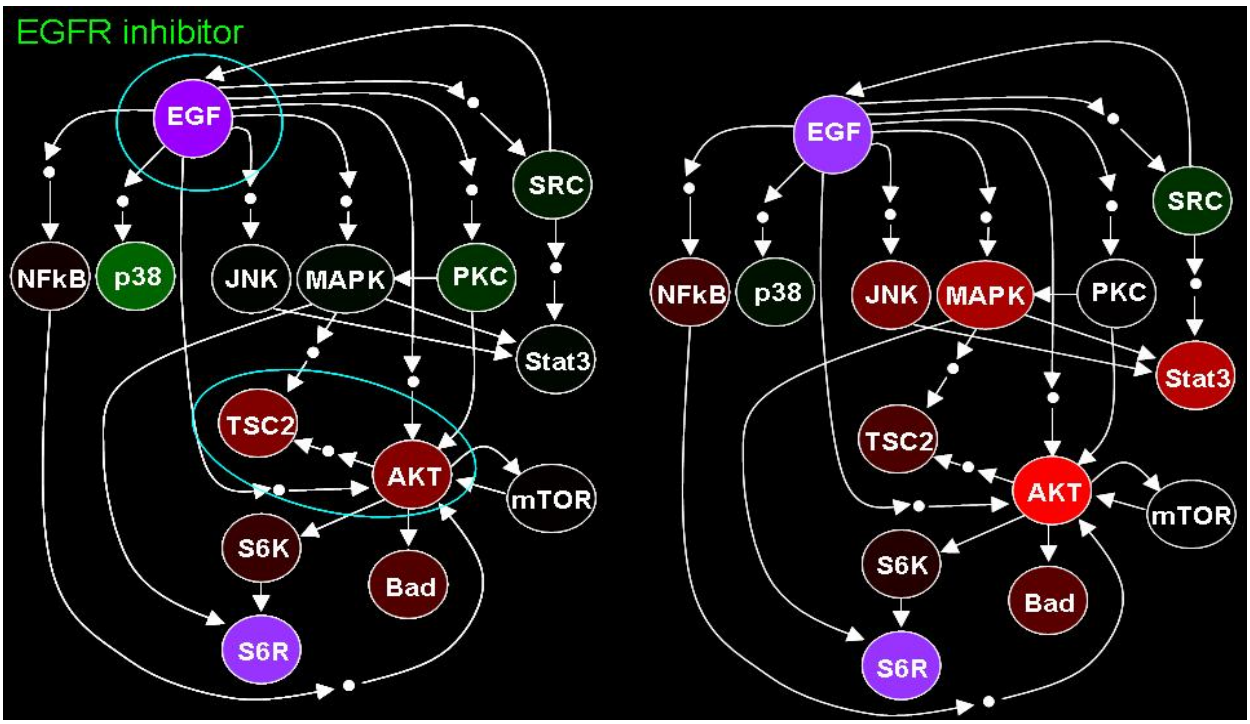

Figure 3. Network changes upon addition of Iressa (EGFR inhibtor).

We next determined the changes in the network upon targeted manipulations of the signaling network. We have used siRNA as well as targeted pharmacological agents. Data shown in figure 4 is from the breast tumor cells treated with PD98059 (MEK inhibitor) and perifosine (AKT inhibitor).

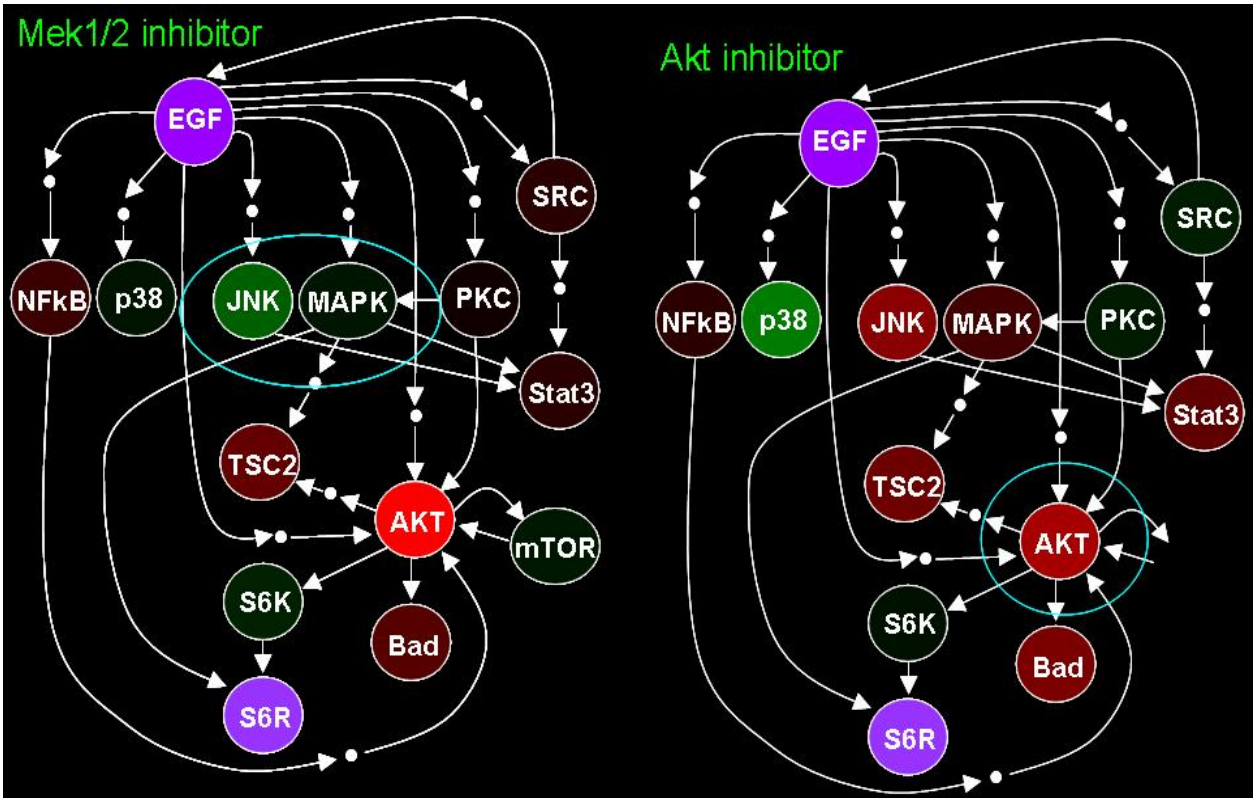

Figure 4. Changes in information flow upon targeted manipulations.

From this data we observe the expected decrease in MAPK activity in response to MEK inhibitor (right panel) and the decrease in AKT with the AKT inhibitor (left panel). We have incorporated these data into our network model to enable us to tackle Task 3 which is to determine combinations of molecules to target.

Data from Task 3 is seen on the following page.

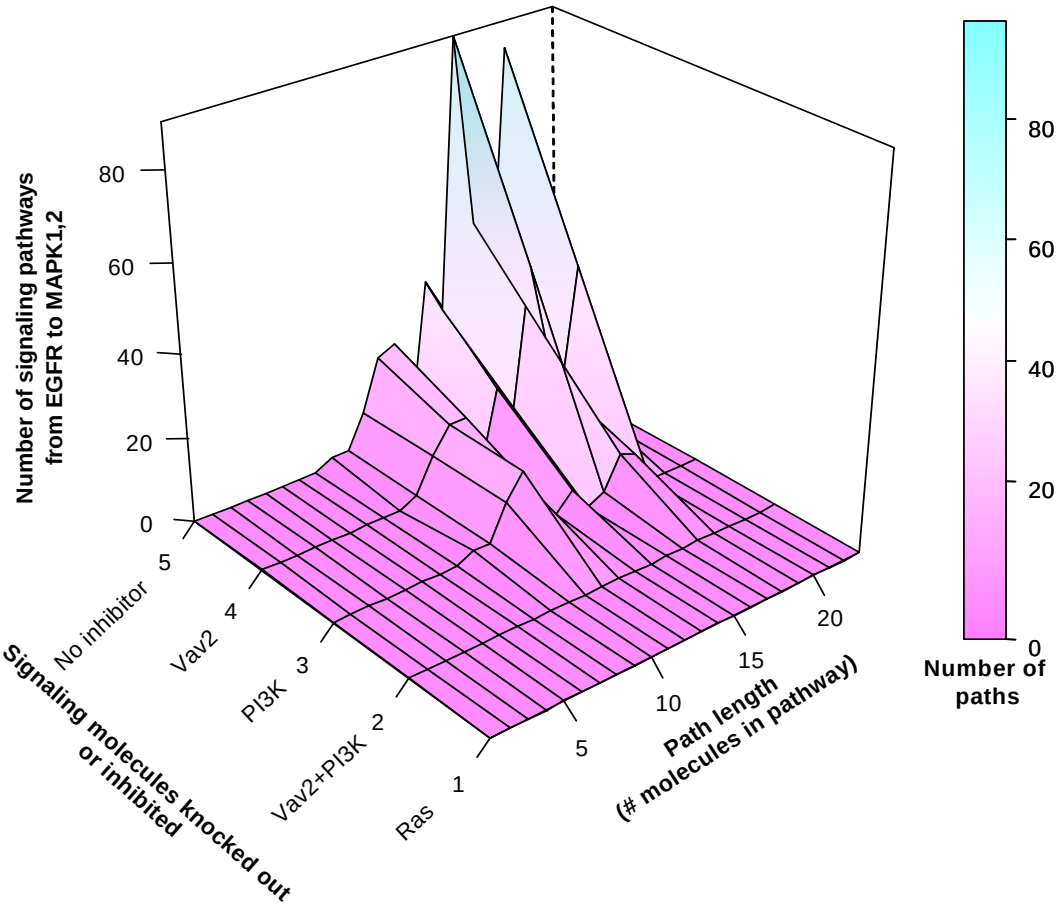

Figure 5 Network analysis of connectivity from EGFR to MAPK1,2. Y-axis, number of possible signaling pathways. X-axis targeted inhibition of signaling molecules. Z-axis path length or molecules in the signaling pathway.

The modeling results suggest that if Ras is inhibited there is a complete loss of connectivity from EGFR to MAPK1,2. If either Vav2 or PI3K are inhibited individually there is a partial loss in connectivity. Interestingly, if both PI3K and Vav2 are knocked there is complete loss of **c**onnectivity, suggesting that a Vav2 and PI3K inhibitor combination can block MAPK1,2 activation

The data presented here pertain to task 1C, and 2B which are completed, and Task 3 which is currently ongoing.

Key research accomplishments

Task 1C. We have determined the network response to EGF inhibitors.

Task 2B. We have determined the dynamic changes in activity of the network in response to growth factor stimulation and are integrating this data into the computational model.

Task 3. We have developed and made predictions from our model to identify combination of targets.

Reportable outcomes
From the work that we have done in the past year we have one manuscripts accepted and another manuscript in review.

Conclusions
We have developed a computational model of the signaling network. We have also generated data of the dynamic changes in signaling and are currently integrating the biological data with the computational model. We are in the final aspects of the proposal and are testing the predictions from our model to identify novel combinations.

Reference (papers from our work published/submited this year)
Ruths D, Tseng J-T, Nakleh L, Ram PT DeNovo Signaling Pathway Predictions based on protein-protein interaction, targeted therapy and protein microarray analysis. *Systems Biology and Computational Proteomics* 2007 109-119 In Press

Muller M, Obeyesekere M, Mills GM, Ram PT Network topology determines dynamics of the mammalian MAPK1,2 signaling network: bi-fan motif regulation of C-Raf and B-Raf isoforms by FGFR and MC1R. *Submitted*

Appendix
Ruths D, Tseng J-T, Nakleh L, Ram PT DeNovo Signaling Pathway Predictions based on protein-protein interaction, targeted therapy and protein microarray analysis. *Systems Biology and Computational Proteomics* 2007 109-119 In Press

De Novo Signaling Pathway Predictions Based on Protein-Protein Interaction, Targeted Therapy and Protein Microarray Analysis

Derek Ruths[1,⋆], Jen-Te Tseng[2], Luay Nakhleh[1], and Prahlad T. Ram[2]

[1] Department of Computer Science, Rice University, Houston, TX 77005, USA
[2] University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA
druths@cs.rice.edu

Abstract. Mapping intra-cellular signaling networks is a critical step in developing an understanding of and treatments for many devastating diseases. The predominant ways of discovering pathways in these networks are knockout and pharmacological inhibition experiments. However, experimental evidence for new pathways can be difficult to explain within existing maps of signaling networks.

In this paper, we present a novel computational method that integrates pharmacological intervention experiments with protein interaction data in order to predict new signaling pathways that explain unexpected experimental results. Biologists can use these hypotheses to design experiments to further elucidate underlying signaling mechanisms or to directly augment an existing signaling network model.

When applied to experimental results from human breast cancer cells targeting the epidermal growth factor receptor (EGFR) network, our method proposes several new, biologically-viable pathways that explain the evidence for a new signaling pathway. These results demonstrate that the method has potential for aiding biologists in generating hypothetical pathways to explain experimental findings.

Our method is implemented as part of the PathwayOracle toolkit and is available from the authors upon request.

1 Introduction

Altered cellular signaling networks can give rise to the oncogenic properties of cancer cells [8], increase a person's susceptibility to heart disease [6], and are responsible for many other devastating diseases [8,3]. As a result, major efforts are currently underway to establish high-resolution maps of signaling networks for various disease-causing cells. These can be used to inform the development of diagnostic methods and pharmacological treatments.

In the laboratory, targeted manipulation experiments either using knockouts (i.e., siRNA or genetic knockout organisms) or pharmacological agents are a primary method for uncovering new connectivity or parts of a signaling network. The goal of such experiments is to generate results that cannot be explained using existing signaling pathway maps or models. These results are important because they signal the discovery of new pathways, but at the same time raise the very open-ended issue of identifying the cause of the incongruous result.

⋆ Corresponding author.

T. Ideker and V. Bafna (Eds.): Syst. Biol. and Comput. Proteomics Ws, LNBI 4532, pp. 109–119, 2007.

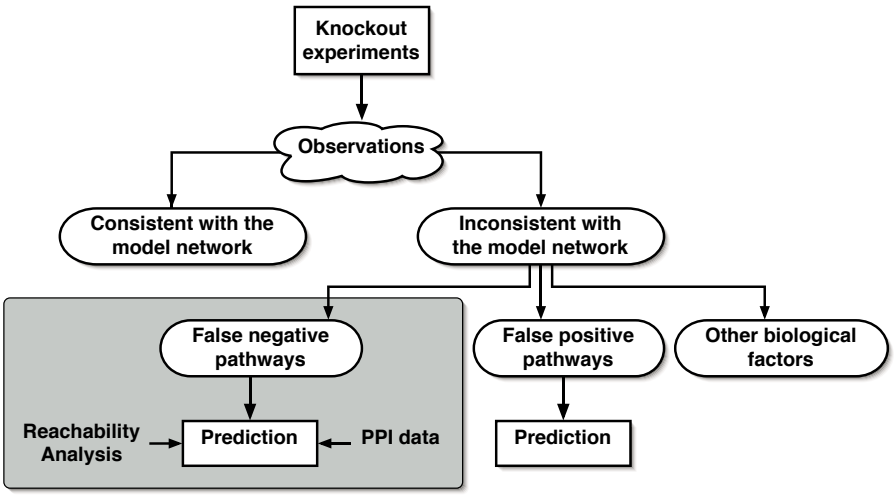

Fig. 1. The path from experiment to new biological insights. Informative knockout or inhibition results are those that cannot be explained by the model. Once such a result has been obtained, the biologist must consider the possible causes for the inconsistency. This paper handles the case of an incomplete signaling model (in the grey box) by providing a computational method for detecting absent pathways and predicting new ones.

As shown in Fig. 1, several explanations can account for unexpected results:

1. *The model is missing signaling pathways.* In this situation, the result is unexpected because interaction paths exist in the biological signaling network that are not represented in the model. These missing paths are false negatives since the model indicates that no such paths exist.
2. *The model contains incorrect signaling interactions or pathways.* Particularly when dealing with diseased cells, signaling network models based on different cell lines can be inaccurate: interactions in one cell line may not exist in the diseased network under study. Thus, the model contains paths that are false positives—paths that do not exist in the context of the cell being studied.
3. *Biological factors have influenced the result.* These can range from technical challenges such as experimental conditions to issues of great scientific importance such as a lack of specificity in the drug being used to knockout or inhibit part of the network.

Thus, when faced with an unexpected result from a knockout or inhibition experiment, the biologist has a large space of potential causes that he or she must consider. As a result, there is a significant need to develop tools that expedite the process of generating hypotheses to explain unexpected targeted manipulation experimental results.

In this paper, we present a novel computational method for identifying and handling knockout or inhibition results that belong to the first class discussed above—those that cannot be explained because the model is missing pathways. Our method (1) identifies results for which the model network is missing paths and (2) generates biologically-viable

pathways that can explain the result. These generated pathways become hypotheses that the biologist can then use as a basis for further experiments or as paths that are added to the existing network model. Prior work in this area has focused on related problems in the transcriptional network domain [20,21]. However, to our knowledge, this method is the first to use knockout or inhibition experiments to guide the prediction of missing pathways in the cellular signaling network.

In order to generate new pathways, our approach integrates knockout or inhibition result data with protein interaction data—both sources of information about interactions that occur in signaling networks.

In a knockout or inhibition experiment, one or more compounds in the signaling network are rendered inactive through chemical or genetic means. In the resulting network, any role that these compounds played are eliminated. The modified network is stimulated and set into motion. At various time intervals, the concentration and activity of various proteins within the modified network are compared to those in the original network. A statistically significant change in the concentration or activity of a given protein in the modified network indicates that this protein and the inhibition target must interact. A reasonable representation of such a positive result is the knowledge that a protein X interacts with another protein Y. Since this captures the interaction information supplied by the experiment, this is the representation we use throughout this paper.

Protein interaction data, commonly stored in protein-protein interaction databases, is another major source of interaction information. This data is primarily generated by high-throughput experimental methods that identify protein pairs that are likely to interact. Unlike the results of knockout or inhibition experiments, all interactions returned by these high-throughput methods are putative. As a result, the false positive rate in protein interaction databases has been shown to be high [15]. Various methods, ranging from literature search to comparisons across organisms, have been proposed for assessing the likelihood of an interaction being correct [9,4,2,18,16]. When a protein interaction database is coupled with an interaction confidence measure, it becomes a useful source of information on interactions that occur within the cell.

Since signaling networks ultimately are massive webs of directed protein interactions, one might expect that new signaling topology could be uncovered by dissecting these protein interaction databases. Yeang *et al.* considered this question with respect to transcriptional networks [20]. In a more recent study, Scott *et al.* [15] considered this problem with respect to signaling networks and found that highly biologically-relevant topologies could be extracted from these interaction networks. In their analysis, they recovered the MAP kinase and ubiquitin-ligation signaling pathways from a computational search of the MIPS interaction database [12].

Our approach uses this idea of discovering topological structure within a protein interaction dataset by considering it within the context of a single knockout or inhibition experiment. The computational technique searches a protein interaction network for biologically-viable pathways that account for the results of the experiment. We make the assumption that interactions with a high likelihood of being correct are biologically-viable. Extending this assumption to the pathway-level, we consider a pathway to be biologically-viable if the product of the likelihoods of each interaction in the pathway

is high. Therefore, our method searches a protein interaction network for the best supported interaction paths that connect X and Y.

In order to test our method, we experimentally and computationally determined the effect of pharmacological inhibitors on changes in signaling network function in human breast cancer cells. Two human breast cancer cell lines were treated with three different pharmacological inhibitors targeting different signaling molecules. We found an unexpected inhibitory interaction between *MEK1* and *c-Src*. Given this result, our method generates excellent candidate pathways that explain the observed knockout or inhibition pattern and are consistent with other biologically known properties of the *EGFR* network. This result can be taken as evidence that our method's generated pathways can be considered reasonable hypotheses for the true signaling network topology underlying experimental results.

In order to make our method available for use, we have implemented it as a Java tool and bundled it with the PathwayOracle software package. PathwayOracle is available upon request from the authors.

2 Results and Discussion

2.1 Experimental Results

In order to understand how targeted manipulations alter different nodes in the signaling network we used inhibitors to specific molecules and measured changes in several proteins within the network using protein microarrays. Combining targeted pharmacological manipulations with protein array technology allows us to simultaneously measure changes in a large number of signaling molecules very rapidly. Using this method we treated breast cancer calls with three inhibitors of the signaling network.

The inhibitors used were Iressa (EGFR kinase inhibitor), perifosine (AKT inhibitor) and PD98059 (MEK inhibitor). Iressa is currently used in clinical treatment of patients, and AKT and MEK inhibitors are in pre-clinical and early phase clinical trials [7].

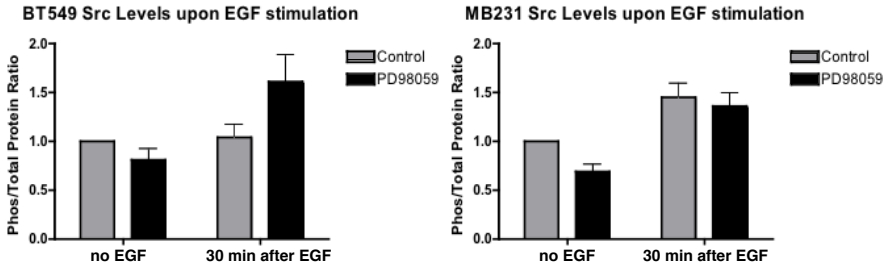

Fig. 2. Experimental microarray data from BT549 and MDA-MB-231 breast tumor cells treated with the MEK1 inhibitor PD98059 shows that the level of phopho c-Src is increased in BT549 cells but not in MDA-231 cells upon EGF stimulation. The two graphs show the phospho c-Src levels in the two cell lines after normalization for protein loading, the first bar corresponds to control cells and the second bar corresponds to cells treated with the MEK1 inhibitor for 30 minutes.

Analysis of the data from the two cell lines at two different time points in which post stimulation revealed changes in signaling within the network (see Figure 2). We observed the expected changes (not shown), i.e. when the MEK inhibitor was used EGF did not stimulate MAPK1,2 but the activation of AKT was not altered. When Iressa was used to inhibit EGFR the activation of MAPK 1,2, was blocked in response to EGF in Ras wild type cells but not in cells with a Ras activation mutation. Similarly Iressa blocked AKT activation of PTEN wild type cells but not in PTEN deletion cells. Having observed expected outcomes we were very intrigued by results that were unexpected. For example we found that in BT549 breast tumor cells PD 98059 elevated c-Src basal phosphorylation levels in EGF stimulated cells. However, this was not the case in MDA-MB-231 cells, where there was no increase in c-Src phosphorylation when compared to control. This data suggests that by inhibiting MEK1 we are also increasing c-Src. There could be two explanations for this result, the first being that MEK and c-Src are connected through a signaling pathway in BT549 cells, or the second being that the MEK inhibitor has non-specific activity on c-Src. However, based on the result in MDA-231 cells where there is no increase in c-Src it does not appear that there is a non-specific drug effect on c-Src. From these results we checked our existing signaling network model to find connectivity between MEK1 and c-Src, and found no existing pathway.

2.2 Pathway Prediction Results

From our experimental data we observe that inhibiting MEK1 results in a 90% decrease in phosphorylation of c-Src in BT549 cells. In order to understand how inhibiting MEK1 could also inhibit c-Src we performed a PubMed search and found no previously published work describing MEK1 activation of c-Src. There were several publications showing that c-Src could activate MEK1, but not vice versa.

Ordinarily when faced with this scenario of having an unexplained experimental outcome and no previously described pathway from MEK1 to c-Src, the biological investigator is faced with hours of literature searches in an attempt to find pair-wise interactions that can connect MEK1 to c-Src. These searches frequently result in several possible best guess pathways that the investigator would then have to check individually. This method of going down a laundry list of pathways to test is very inefficient and uses valuable time, manpower and resources. Computational methods to identify possible pathways focus this effort and allow the investigator to logically rank and test the pathways based on the modeling prediction. We have developed such a method and show here the use of our model and the use of iterative cycling between experiments and modeling to rapidly advance our understanding of signaling networks.

The computational model predicts several pathways from MEK1 to c-Src based on protein-protein interaction data (see Fig. 3). Some of the biologically-relevant characteristics of the predictions include the prediction that all paths include SEK1 and p38 which have been shown to be downstream from MEK1 [17,10]. The fact that our method identified this biologically correct connectivity increases the confidence in the predicted pathways. Downstream from p38 there is a predicted bifurcation of signal with seven possible paths. However, these seven paths converge onto three molecules c-CBL, Caveolin1, and FADK1 which are directly upstream from c-Src.

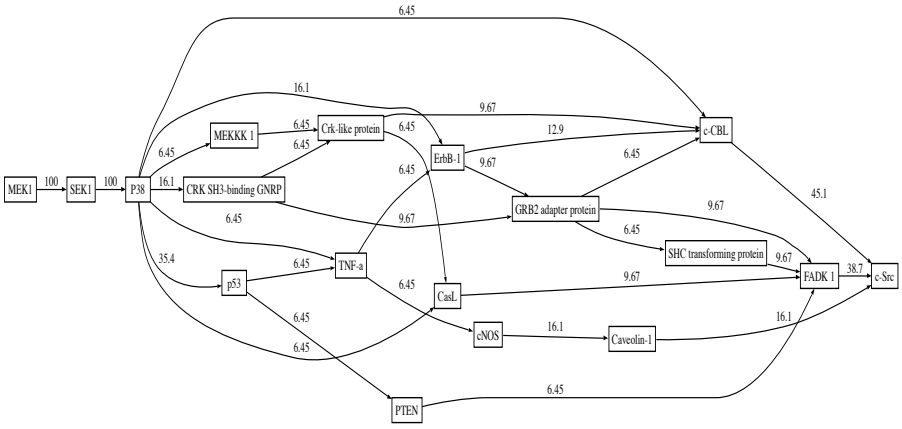

Fig. 3. A graphical representation of the paths predicted leading from MEK1 to c-Src. Each interaction (edges) is labeled by the % of paths that it appeared in. Since this is the percent of predicted paths that pass through a given interaction, this number can be taken as an estimate of the importance of the interaction among all the interactions in the prediction. Note that this number should not be confused with the confidence that the interaction exists—all interactions depicted in this graph had support values greater than 99.9% as reported by the STRING database.

This modeling result is very interesting because it offers testable hypotheses to direct the experimental validation of the predictions. The first experiment is to knock out SEK1 or p38, anticipating that this would completely knock out connectivity between MEK1 and c-Src. Experiments to inhibit the connectivity in this pathway would include using siRNA to knock out expression of SEK1 and p38, and chemical intervention experiment by using a pharmacological inhibitor of p38. If we experimentally observe that, when p38 is inhibited, there is no change in connectivity between MEK1 and c-Src this would direct us back to make changes in the model. If we observe only partial loss of connectivity when p38 is blocked, this would imply additional pathways not utilizing p38, and this again would direct us back to refine our model. Additionally, knocking out or pharmacologically inhibiting c-CBL, Caveolin1, or FADK1 should give one of three results complete, partial, or no loss of connectivity between MEK and c-Src. Based on the results from these experiments we would be able to determine novel connectivity between MEK1 and c-Src in a quick and directed manner. Therefore, by this modeling-based hypothesis-driven method, coupled with targeted experimental manipulations, we can rapidly identify novel connectivity between signaling molecules and pathways.

3 Materials and Methods

3.1 Knockout Experiment Design

In order to quantify changes in several nodes of the signaling network in parallel we used the reverse phase protein micro-array technology. Using this proteomic tool we were able to measure changes in the activity state as well as total levels of expressed proteins. The method is described below.

Protein Lysate Micro Array. Arrays were prepared using cells lysed on ice with microarray lysis buffer (50 mM Hepes, 150 mM NaCl, 1mM EGTA, 10 mM Sodium Pyrophosphate, pH 7.4, 100 nM NaF, 1.5 mM MgCl2, 10% glycerol, 1% Triton X-100 plus protease inhibitors; aprotinin, bestatin, leupeptin, E-64, and pepstatin A). Cell lysates were centrifuged at 15,000 g for 10 minutes at 4C. Supernatant was collected and quantified using using a protein-assay system (Bio-Rad, Hercules, CA), with BSA as a standard. Using a GeneTac G3 DNA arrayer (Genomic Solutions, Ann Arbor, MI), six two-fold serial dilutions of cell lysates are arrayed on multiple nitrocellulose-coated glass slides (FAST Slides, Whatman Schleicher & Schuell, Keene, N.H). Arrays were produced in batches of 10. Printed slides were stored in dessicant at -20C. Antibodies were screened for specificity by Western blotting. An antibody was accepted only if it produced a single predominant band at the expected molecular weight. Each array was incubated with specific primary antibody, which was detected by using the catalyzed signal amplification (CSA) system (DAKO). Briefly, each slide was washed in a mild stripping solution of Re- Blot Plus (Chemicon International, Temecula, CA) then blocked with I- block (Tropix, Bedford, MA) for at least 30 minutes. Following the DAKO universal staining system, slides were then incubated with hydrogen peroxide, followed by Avidin for 5 minutes, and Biotin for 5 minutes. Slides were incubated with primary and secondary antibodies then incubated with streptavidin-peroxidase for 15 minutes, biotinyl tyramide (for amplification) for 15 minutes, and 3,3-diaminobenzidine tetrahydrochloride chromogen for 5 minutes. Between steps, the slide was washed with TBS-T buffer. Each slide was probed with validated antibodies under optimal blocking and binding conditions. Loading is determined by comparing phosphorylated and non-phosphorylated antibodies as well as by assessing control antibodies to prevalent and stable proteins. Six serial dilutions of each sample facilitate quantification and ensure that any slide can be assessed with different antibodies. Multiple controls are placed on each slide to facilitate quantification and robustness of the assay. Data are collected and analyzed by background correction and spot intensity using Image J. Protein phosphorylation levels are expressed as a ratio to equivalent total proteins. Fold increases in spot intensities were calculated against non-stimulated control samples. The following antibodies were used: EGFR, c-Src, Stat3, MAPK1,2, AKT, S6K, MEK1, NFkB, BAD, p38 MAPK, phosho c-Src, phospho Stat3, phospho AKT, phospho S6K, phospho MEK1, phospho NFkB, phospho BAD, phospho p38 MAPK.

3.2 Predicting Novel Pathways Based on Knockout Results

After completing the set of knockout experiments, we conducted a novel computational analysis to predict new pathways needed to explain the experimental results. This analysis consisted of two main stages:

1. *Identifying inconsistent results:* in this step we identified any individual knockout experiments that could not be explained by the model network. We call these results *inconsistent*.
2. *Constructing candidate pathways:* for each inconsistent result, we performed an exhaustive search of protein interaction data for hypothetical pathways that could explain the result and augment the existing incomplete model.

It is important to recall from Fig. 1 that there are multiple explanations for inconsistent results—only one of which is the incompleteness of the model. To be concrete, the experimental results presented in this paper can also be explained by undesired drug interactions with proteins other than MEK1. Our analysis finds several very viable pathways that may be missing from this network, making our approach valuable to the experimental biologist. However, in a complete analysis other sources of error must be taken into account. We identify these other sources of inconsistency as directions for future work, focusing in this paper only on the prediction of new pathways to handle the case of an incomplete model.

In the following sections we provide a detailed description of the steps itemized above.

Identifying Inconsistent Results. In order to determine which experimental results were unexpected, it was necessary to select a model signaling network that contained the complete set of known and relevant interactions. Since all of our experiments involved proteins embedded in the EGFR network, we used a model based on an extensive literature review of interactions in this network [11]. We stored the model signaling network as a pathway graph model [14]. In this representation, each protein/protein-state pair (e.g. AKT-inactive, AKT-active, and EGFR-phosphorylated) and each interaction is represented by a node. Directed edges connect protein/state pairs to interactions (reactions) they participate in and connect reactions to protein/state pairs that are produced as a result of the interaction. This representation explicitly depicts all experimentally derived and published paths through the signaling network—allowing extensive analysis of the connectivity within the network.

Recall that a knockout or inhibition result can indicate that a signaling pathway exists between two proteins (as was the case with *MEK1* and *c-Src* in the experiments described above). When a knockout or inhibition experiment yields such a result for proteins X and Y, but no chain of directed interactions exists in the model network between X and Y, we call this result *inconsistent*—implying that the model is not capable of explaining the result and requires the addition of a new pathway.

In order to identify inconsistent results, we first selected only those results which indicated the presence of a signaling pathway between two proteins. For each of these results, we used the constrained downstream algorithm [14] to enumerate all paths between the two proteins in the model. This algorithm performs an exhaustive search of a pathway graph model for all paths connecting one set of proteins to another. In this algorithm, the first protein is considered the source, the second protein is considered the sink, and all paths found are directed from the sources to sinks, as they would occur in the signaling network.

For the experiments we considered for this paper, the downstream algorithm reported paths for all results except *MEK1* to *c-Src*. The absence of any path from *MEK1* to *c-Src* indicates that the model cannot explain the inhibitory result observed between these two proteins. As a result, this result was identified as an inconsistent result.

Constructing Candidate Pathways. In this step, given an inconsistent result, we seek a set of candidate pathways, any of which can explain the result observed. For the inconsistent result supporting a pathway between proteins X and Y, we know that the

model has insufficient interactions to connect them. Therefore, we must look elsewhere in order to find biologically-relevant interactions to connect these two proteins.

Protein interaction databases are, effectively, massive repositories of putative protein interactions. Despite the fact that many of the interactions may not, in reality, occur, these databases provide a good source of interactions to use when assembling hypothetical pathways.

One issue that must be addressed is the fact that many studies have shown the interactions in these databases to be of varying quality [4,2]. Since we seek biologically-likely pathways which are, by definition, composed of biologically likely interactions, we must have some way of evaluating the *confidence* of any given interaction in the database. Significant work has been done into the problem of assigning confidence to interactions [9,4,2,18,16]. In this study, we made use of the STRING database [19] which provides interactions with confidence scores. However, using other interaction databases and other confidence scoring schemes are equally valid approaches and, depending on the interactions in the database and how confidence is estimated, may produce somewhat different results from ours.

Once a protein interaction database and confidence scoring scheme have been selected, a protein interaction network can be constructed. This is a data structure that combines the interactions in the database with the scoring scheme. In this network, a node is a protein, an edge $e = (u, v)$ is an undirected interaction between proteins u and v. Each edge, $e = (u, v)$ is assigned a weight equal to its log-likelihood score: $weight(e) = -log(c(e))$, where $c(e)$ is the confidence assigned to interaction e by the scoring scheme.

When constructed as described, this network has the special property that the weight of path $\langle u_1, u_2, ..., u_n \rangle$ within this network has the following correspondence to its total support:

$$\sum_{i=1}^{n-1} w((u_i, u_{i+1})) = -log(\prod_{i=1}^{n-1} c((u_i, u_{i+1}))).$$

Since the function $-log(x)$ approaches 0 as $x \rightarrow 1$, the sum on the left will be smallest when the individual path edges have confidence scores closest to 1. Therefore, the shortest (lightest) path in the network between nodes X and Y corresponds to the most biologically-likely pathway connecting the two proteins represented by nodes X and Y.

Since all paths within some confidence threshold probably correspond to some biologically-likely pathway, we choose to search for the set of k-shortest paths—where k is a parameter indicating how many paths we want to retrieve. Paths should be reported in order of increasing weight so that the kth path is the longest (least biologically-likely) of the paths returned by the search.

Significant work has been done on the problem of enumerating the k-shortest paths and efficient algorithms exist for solving it [5,1]. For our purposes in this project, we use a variant of the k-shortest path problem, called the k-shortest *simple* path problem [22,13]. A simple path is one that contains no loops. The reason for this restriction is that, while feedback loops are quite common in signaling pathways, we are only interested in the simplest pathways that can explain the inconsistent results. Under the log-likelihood transformation, edges with 100% support will have zero weight, creating

the possibility of cycles in the graph. As a result, we choose to discard any short paths that contain loops from the set of candidate pathways.

In our analysis, we used an implementation of Eppstein's k-shortest paths algorithm [5]. Non-simple paths were detected and removed from the output in order to give a k-shortest simple paths algorithm. We ran the algorithm and found the 100 shortest simple paths. A detailed analysis of these paths is given in Section 2.2.

As a final step in identifying the candidate pathways, direction must be imposed on the paths extracted. The paths extracted from the protein interaction network are bi-directional since the edges are undirected. For a result in which a knocking out protein X caused a change in protein Y, the pathway direction is towards protein Y. In order to model this in the interaction network, we always search for paths from X to Y and report the the nodes of each path in the order in which they appear—from first to last.

3.3 The PathwayOracle Tool

In the past ten to fifteen years biologists have uncovered hundreds of interactions within signaling pathways in biological systems. A challenge given this large amount of data is to develop novel methods to probe the data and ask questions that cannot be answered by experimental biology alone. On the other hand it is also vital to integrate the experimental biology with the computational models and methods.

In order to address these issues, we have created the PathwayOracle software package which contains various tools enabling the computational analysis and extension of experimental results and techniques [14]. The novel approach to pathway prediction described in this paper is the most recent addition to the PathwayOracle package. Included with the implementation is the human subset of the interactions in the STRING database, though other interaction datasets can be specified.

The entire toolkit is open-source, implemented in Java, and available upon request from the authors. Additional information about other features and tools included in the package is available on the website:
`http://bioinfo.cs.rice.edu/pathwayoracle`.

Acknowledgments

We gratefully acknowledge the help that Melissa Muller provided in assembling the figures for the experimental results. We appreciate Erion Plaku's time and effort spent debugging an implementation of the shortest path algorithm.

This work was supported in part by DOD grant BC044268 to PTR.

References

1. Ahuja, R., Mehlhorn, K., Tarjan, B.: Faster Algorithms for the Shortest Path Problem. Journal of Association of Computing Machinery 37, 213–223 (1990)
2. Bader, J., Chaudhuri, A., Rothberg, J., Chant, J.: Gaining confidence in high-throughput protein interaction networks. Nature Biotechnology 22(1), 78–85 (2004)
3. Belloni, E., Muenke, M., Roessier, E., Traverse, G., Siegel-Bartelt, J., Frumkin, A., Mitchell, H.F., Donis-Keller, H., Helms, C., Hing, A.V., Heng, H.H.Q., Koop, B., Martindale, D., Rommens, J.M., Tsui, L.-C., Scherer, S.W.: Identification of Sonic hedgehog as a candidate gene responsible for holopro-sencephaly. Nature Genetics 14, 353–356 (1996)

4. Deng, M., Sun, F., Chen, T.: Assessment of the reliability of protein-protein interactions and protein function prediction. In: Proceedings of the Eight Pacific Symposium on Biocomputing, pp. 140–151 (2003)
5. Eppstein, D.: Finding the k shortest paths. SIAM Journal of Computing 28(2), 652–673 (1998)
6. Feldman, D.S., Carnes, C.A., Abraham, W.T., Bristow, M.R.: Mechanisms of Disease: β-adrenergic receptors alterations in signal transduction and pharmacogenomics in heart failure. Nature Clinical Practice Cardiovascular Medicine 2, 475–483 (2005)
7. Hennessy, B.T., Smith, D.L., Ram, P.T., Lu, Y., Mills, G.B.: Exploiting the PI3K-AKT pathway for cancer drug discovery. Nature Review Drug Discovery 4, 988–1004 (2005)
8. Hunter, T.: Signaling – 2000 and beyond. Cell 100(1), 113–127 (2000)
9. Hwang, D., Rust, A.G., Ramsey, S., Smith, J.J., Leslie, D.M., Weston, A.D., Atauri, P., Aitchison, J.D., Hood, L., Siegel, A.F., Bolouri, H.: A data integration methodology for systems biology. PNAS 102(48), 17296–17301 (2005)
10. Johnson, G.L., Lapadat, R.: Mitogen-activated protein kinase pathways mediated by ERK, JNK, and protein kinases. Science 1912, 38 (2002)
11. Oda, K., Matsuoka, Y., Funahashi, A., Kitano, H.: A comprehensive pathway map of epidermal growth factor signaling. Molecular Systems Biology, msb41000014–E1–E17 (2005)
12. Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kalthenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H., Reupp, A., Frishman, D.: The MIPS mammalian protein-protein interaction database. Bioinformatics 21(6), 832–834 (2005)
13. Pascoal, M., Martins, E.: A new implementation of Yen's ranking loopless paths algorithm. Quarterly Journal of the Belgian, French, and Italian Operations Research Societies 1(2), 121–134 (2003)
14. Ruths, D., Nakhleh, L., Iyengar, M.S., Reddy, S.A.G., Ram, P.T.: Graph-theoretic Hypothesis Generation in Biological Signaling Networks. Journal of Computational Biology. (In press) (2006)
15. Scott, J., Ideker, T., Karp, R.M., Sharan, R.: Efficient Algorithms for Detecting Signaling Pathways in Protein Interaction Networks. In: McLysaght, A., Huson, D.H. (eds.) RECOMB 2005. LNCS (LNBI), vol. 3678, pp. 1–13. Springer, Heidelberg (2005)
16. Sharan, R., Suthram, S., Kelly, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T.: Conserved patterns of protein interaction in multiple species. PNAS 102(6), 1974–1979 (2005)
17. Uhlik, M.T., Abell, A.N., Cuevas, B.D., Nakamura, K., Johnson, G.L.: Wiring diagrams of MAPK regulation by MEKK1, 2, and 3. Biochem. Cell Biol. 82(6), 658–663 (2004)
18. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399–403 (2002)
19. von Mering, C., Heynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B.: STRING: a database of predicted functional associations between proteins. Nucleic Acids Research 31(1), 258–261 (2003)
20. Yeang, C.H., Ideker, T., Jaakkola, T.: Physical network models. Journal of Computational Biology 11, 243–262 (2004)
21. Yeang, C.H., Mak, H.C., McCuine, S., Workman, C., Jaakkola, T., Ideker, T.: Validation and refinement of gene regulatory pathways on a network of physical interactions. Genome Biology 6(7), R62 (2005)
22. Yen, J.Y.: Finding the K Shortest Loopless Paths in a Network. Management Science 17(11), 712–716 (1971)